The Evolution of the Phage Shock Protein Response System: Interplay between Protein Function, Genomic Organization, and System Function

M. Huvet,*[,1] T. Toni,[1,2] X. Sheng,[1] T. Thorne,[1,2] G. Jovanovic,[3] C. Engl,[3] M. Buck,[3] J.W. Pinney,*[,1] and M.P.H. Stumpf*[,1,2,4]

[1]Centre for Bioinformatics, Division of Molecular Biosciences, Imperial College London, London, United Kingdom

[2]Institute of Mathematical Sciences, Imperial College London, London, United Kingdom

[3]Division of Biology, Imperial College London, London, United Kingdom

[4]Centre for Integrative Systems Biology, Imperial College London, London, United Kingdom

***Corresponding author:** E-mail: m.huvet@imperial.ac.uk; j.pinney@imperial.ac.uk; m.stumpf@imperial.ac.uk.

Associate editor: Oliver Pybus

Research article

Abstract

Sensing the environment and responding appropriately to it are key capabilities for the survival of an organism. All extant organisms must have evolved suitable sensors, signaling systems, and response mechanisms allowing them to survive under the conditions they are likely to encounter. Here, we investigate in detail the evolutionary history of one such system: The phage shock protein (Psp) stress response system is an important part of the stress response machinery in many bacteria, including *Escherichia coli K12*.

Here, we use a systematic analysis of the genes that make up and regulate the Psp system in *E. coli* in order to elucidate the evolutionary history of the system. We compare gene sharing, sequence evolution, and conservation of protein-coding as well as noncoding DNA sequences and link these to comparative analyses of genome/operon organization across 698 bacterial genomes. Finally, we evaluate experimentally the biological advantage/disadvantage of a simplified version of the Psp system under different oxygen-related environments.

Our results suggest that the Psp system evolved around a core response mechanism by gradually co-opting genes into the system to provide more nuanced sensory, signaling, and effector functionalities. We find that recruitment of new genes into the response machinery is closely linked to incorporation of these genes into a *psp* operon as is seen in *E. coli*, which contains the bulk of genes involved in the response. The organization of this operon allows for surprising levels of additional transcriptional control and flexibility. The results discussed here suggest that the components of such signaling systems will only be evolutionarily conserved if the overall functionality of the system can be maintained.

Key words: comparative genomics, bacteria, signaling system, phage shock stress response, signal transduction network.

Introduction

The ability to process and respond to environmental signals is an essential feature of living organisms. A change in the environment, for example, when a microbial pathogen enters its host, can have profound effects on the internal state and dynamics. Typically, stress response mechanisms are then initiated in order to adapt to the new surroundings, which may include changes in oxygen availability, ambient pH, osmolarity, salt concentration, nutrient availability, or attack from the innate and adaptive immune systems. Maintaining suitable response mechanisms is thus a necessary prerequisite for pathogens to establish successful infections and persistence inside susceptible hosts and more generally for survival in changing and often hostile environments. Here, we investigate how the molecular machinery underlying one such response was assembled and has evolved in bacteria.

In *Escherichia coli*, the phage shock protein (Psp) stress response system is responsible for repairing damage to the inner membrane of the cell (Kobayashi et al. 2007) and maintenance of the proton-motive force (pmf) across the inner membrane (reviewed in Model et al. 1997; Darwin 2005; Jovanovic et al. 2006). The Psp stress response is invoked, for example, during phage infection, secretin production, blockage of protein export or fatty acid/phospholipid biosynthesis, exposure to organic solvents, extreme heat or osmotic shock, high ambient pH, and action of protonophores, and has been characterized in great detail (reviewed in Model et al. 1997; Darwin 2005). The Psp system appears to be widely shared across sequenced *Enterobacteria* (and more generally gram-negative bacteria) (Darwin 2005; Huvet et al. 2009). The Psp system has also been implicated in the virulence of *Salmonella*, *Shigella*, and *Yersinia* (reviewed in Model et al. 1997; Darwin 2005, 2007; Rowley et al. 2006).

The response in *E. coli* is mediated by nine proteins that are organized into five units: the *psp* operon, containing the genes coding for PspA, PspB, PspC, PspD, and PspE; and four single-gene units for PspF, PspG, and the

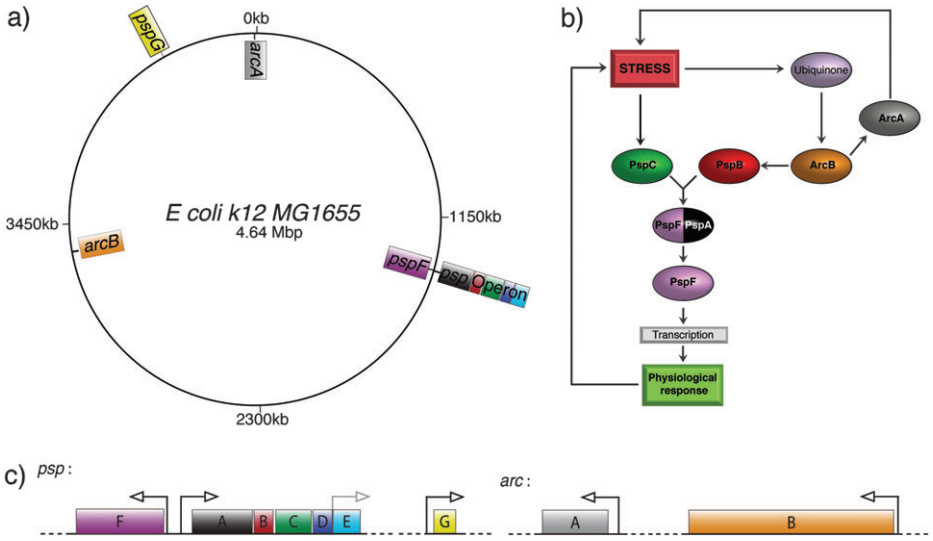

Fig. 1. Organization of the Psp system in *Escherichia coli* K12. (*a*) Position of Psp system genes on the *E. coli* K12 circular chromosome. (*b*) Schematic representation of the Psp system mechanistic model under stress conditions. The ovoid objects represent cellular components and the rectangles biological processes. (*c*) Schematic representation of the transcriptional units making up the Psp system. The arrows represent transcription initiation sites (black: strong and gray, weak). The color code used to designate genes in this figure will be used throughout the remaining figures in the paper (and the supplementary material, Supplementary Material online).

two-component system comprised of ArcA and ArcB (see fig. 1). The *pspABCDE* operon together with *pspF* and *pspG* form the Psp regulon. PspF negatively controls its own expression and positively controls the rest of the regulon by binding at the *pspA* and *pspG* promoters (Model et al. 1997; Lloyd et al. 2004; Darwin 2005); no other specific binding sites for the transcription factor PspF are known or have been predicted in the *E. coli* genome.

Although the physiological role of the system and the functions of all the proteins (PspD, PspE, and PspG) have not yet been fully elucidated, the outline regulatory processes have been well characterized in *E. coli*. Under non-stress conditions, PspA binds and inhibits PspF, forming a PspA–PspF inhibitor complex (Elderkin et al. 2002, 2005; Joly et al. 2009). Under stress conditions, activation of the membrane-bound proteins PspB and PspC, most likely via structural changes in these proteins, mediates the release of PspF from its negative regulation by disrupting the PspA–PspF inhibitory complex. This process initiates, via the binding of PspF to its target upstream DNA-binding sites, the transcriptional response to the stress conditions (Weiner et al. 1991; Gueguen et al. 2009). One main effector of the response appears to be the PspA multimer, which helps to reestablish pmf across the inner membrane (Kobayashi et al. 2007). The nonorthodox two-component signaling (TCS) system consisting of the phosphorelay kinase ArcB and its cognate receptor ArcA influences the strength and effectiveness of the transcriptional response to the inducing stress conditions in microaerobiosis (Jovanovic et al. 2009). However, the ArcB–ArcA system is also known to be involved directly in other stress response systems (Malpica et al. 2006) and so will be considered here in parallel to the genes making up the Psp regulon. Previous data-driven attempts in evolutionary systems biology have focused predominantly on the structure of such networks and how they have changed over (evolutionary) time (Luscombe et al. 2004; Babu et al. 2009). These systems-level evolutionary analyses often investigate the extent of and causes for statistical correlations between evolutionary properties of proteins and their position or role in protein interaction or metabolic networks (Agrafioti et al. 2005; Drummond et al. 2006; Lee et al. 2008; Feist et al. 2009). Here, we focus instead on the function of a particular stress-related signaling network. Adopting this perspective requires us to consider systems that have been experimentally characterized, ideally in several species. For many of the important stress response systems, including osmotic, hypoxia, limitation of combined nitrogen, pH, heat shock, and the phage shock stress response, we do have extensive functional data and, in an increasing number of cases, also mechanistic models with varying degrees of detail and sophistication (Hlavacek et al. 2006).

Here, we perform a systematic and comprehensive analysis of the Psp system by 1) identifying orthologs in 698 fully sequenced bacterial genomes, 2) inferring the patterns of gain and loss of genes over the course of evolutionary history, 3) assessing the rates of sequence divergence and evolution, 4) testing for coevolution among the proteins making up the Psp stress response system, 5) comparing potential Psp stress response systems across different bacterial species, and 6) test experimentally the impact of a simplified Psp system in different oxygen-related environments.

Materials and Methods

Sequences and Annotations

The sequences and annotations for 698 bacterial species considered here were retrieved from the "Genome Assembly/Annotation Projects" data in the NCBI ftp (ftp://

ftp.ncbi.nih.gov/genomes/Bacteria/). For each species, we used the refSeq annotation library.

Ortholog Identification

A reciprocal best hit approach was used to identify orthologs (using Blastall 2.2.19, ftp://ftp.ncbi.nih.gov/blast/executables/LATEST/). Each protein of *E. coli K12 MG1655* was aligned against the full protein set of each of the 697 remaining bacterial species. The first hit of each Blast run was then aligned against the full set of *E. coli K12 MG1655* proteins. Two proteins were defined as potential orthologs if and only if the first hit of the second blast search corresponded to the protein used as a query in the first one (Moreno-Hagelsieb and Latimer 2008).

To minimize the number of false positive, those results were filtered: If the alignment involved less than 50% of the smallest protein and the sequence identity was lower than 30%, the protein pair was excluded from the set of potential orthologs.

In *E. coli APEC 01*, this procedure first did not discover orthologs for PspB and PspD, despite the remaining genes in the *psp* operon showing a genomic organization that was in agreement with the observed one in other *E. coli* species. This is due to erroneous annotation in the published genome sequence as a blastall search revealed homologs (with an almost perfect sequence identity compared with *E. coli K12 MG1655*) to both genes at the expected positions. Following this observation, we decided to apply a blastall search for all *psp* genes in a subset of species used in this study close to *E. coli* to evaluate the proportion of potentially missing, or misannotated, genes. This approach revealed four potential missing genes to add to the two from *E. coli APEC 01*.

Phylogenetic Reconstruction

The species phylogeny was inferred using the 23S rRNA (Ludwig and Schleifer 1994). Usually, each species contains several copies of the 23S rRNA, and to apply a classic method of phylogeny inference, one sequence per organism was chosen at random. On this set of 23S rRNA sequence per organism, a multiple alignment was performed using MAFFT 6.611 (http://align.bmr.kyushu-u.ac.jp/mafft/software/) and used in PhyML 2.4.4 (http://atgc.lirmm.fr/phyml/) to infer the species tree. Different alignment and phylogeny inference procedures resulted in very similar trees. Of the 127 species of interest, just 125 species contain sequences annotated as 23S rRNA leading to a tree of 125 species and not 127.

Coinheritance of Gene Pairs

We used the software BayesTraits (http://www.evolution.reading.ac.uk/BayesTraits.html) to compute the likelihood of two evolutionary models: The first model corresponds to independent evolution of the two genes, whereas the second model assumes that their evolution is correlated (or anticorrelated); here, correlation (anticorrelation) means that the two genes are seen together more (less) often than we would expect given their distribution across the phylogeny. A standard likelihood test ratio was then used to test if the independent model could be rejected in the light of the results obtained for the correlated evolution model. Because of the large number of statistical tests involved, we applied a multiple-testing (false discovery rate [FDR]) correction.

Statistical Control for Ortholog Sharing

The system investigated here contains five genes that in *E. coli* reside in the same operon. Because of this, they are more likely to be inherited together and the operon structure thus forms a potentially confounding factor in all further evolutionary analyses. We therefore studied the genes in all other operons that (in *E. coli*) contain five genes. The operon information was obtained from RegulonDB (http://regulondb.ccg.unam.mx/index.jsp), and orthologs were identified along the same lines as described above. For each operon, the number of species containing 0, 1, 2, 3, 4, and 5 orthologs were then computed (from our set of species used in the detailed analysis of the Psp system's evolution).

We also studied the profile of orthologs for randomly selected set of four orthologs. We extracted 1,000 sets of four proteins not included in operons (according to RegulonDB) and computed (in the species of interest) the number of species containing 0, 1, 2, 3, and 4 orthologs, respectively. Finally, we tested whether the observed number of species containing an ortholog for the four non–operon-based proteins of the Psp system is equivalent to the population of four randomly selected proteins from *E. coli*.

Identification of Extended Operons

The identification of potential extended versions of the *psp* operon was done using intergenic distance and the organization of genes surrounding each ortholog of the *psp* operon genes. Analysis of *E. coli* operons shows that genes in operons tend to be separated by intergenic segments that are shorter than 20 bp, but sometimes segments of 50 bp or more occur (Karimpour-Fard et al. 2008). Therefore, we have used 100 bp as a crude but conservative cutoff to determine whether two genes could be part of the same operon.

Computation of the dN/dS Ratio

The dN/dS ratio was computed with the "kaks" R function from the "seqinr" package (http://cran.r-project.org/web/packages/seqinr/index.html). Multiple alignments were obtained for all sets of orthologous proteins. For each sequence, the amino acids were then replaced by the DNA sequence, maintaining the codon organization. The dN/dS ratio was computed for each ortholog pair using the respective sequences from the multiple alignment. To obtain the percentage identity, the same approach was applied on the amino acid multiple alignment. The computation itself was performed using the "dist.alignment" function (matrix option set to "identity") of "seqinr."

Bacterial Strains and Growth Conditions

The wild-type bacterial strain used in this study was MG1655 (Lloyd et al. 2004). The bacterial strain MVA49

(MG1655*ΔpspBC ΔpspG*::kan, Kan^R) was constructed by transducing a *ΔpspG*::kan (Kan^R) from MVA40 (Lloyd et al. 2004) into MG1655*ΔpspBC* (Lloyd et al. 2004). Strain MVA103 (MVA49) (*pspA-lacZ*), Kan^R Amp^R) was constructed by transducing a ϕ(*pspA-lacZ*) (Amp^R) from MC4100λpsp3 (Dworkin et al. 2000) into MVA49. Transductions were carried out using the $P1_{vir}$ bacteriophage as described in Miller (1992). Strains were routinely grown in Luria-Bertani (LB) broth or on LB agar plates at 37 °C (Miller 1992). For aerobic growth, overnight cultures of cells were diluted 100-fold into 5 ml (or 20 ml for growth rate experiment) of LB in a universal tube with loose-fitting caps and shaken at 200 rpm. For microaerobic growth, overnight cultures of cells were diluted 100-fold into 5 ml (or 20 ml for growth rate experiment) of LB and shaken at 100 rpm. The cells were then taken for either measurements of cell growth rates at OD_{600}, β-gal assays, or western blot analyses. For an anaerobic growth, 25 ml of overnight cultures grown at 37 °C in a universal tube with tightly closed caps without shaking were transferred (100-fold dilution) into a fully LB-filled suba-sealed universal tube (to avoid any air space) and incubated overnight at 37 °C without shaking. The cells were then taken by syringe for either measurements of cell growth rate at OD_{600}, β-gal assays, or western blot analyses. The pIV secretin was constitutively expressed from pGJ4 (Tet^R) (Engl et al. 2009). Antibiotics were routinely used at the following concentrations: ampicillin (Amp; 100 μg ml^{-1}), kanamycin (Kan; 25 μg ml^{-1}), and tetracycline (10 μg ml^{-1}).

The difference of induction of PspA in wild-type and *pspBCG* triple mutant genetic backgrounds was tested with a Mann–Whitney test. We tested the null hypothesis that the induction in the wild-type background is not greater than the one in the *pspBCG* triple mutant. In aerobiosis and microaerobiosis, the null hypothesis is rejected at a 5% level confidence, whereas in anaerobiosis, the null hypothesis cannot be rejected with the same level of confidence.

We then analyzed the growth behavior of the studied strains, using two Kolmogorov–Smirnov and Mann–Whitney tests. First, the impact of pIV on the growth of wild-type and *pspBCG* triple mutant in aerobic and microaerobic condition was tested using the Kolmogorov–Smirnov test. We test whether for a given genotype the growth rates are equivalent in no stress and pIV environments. For both genotypes, the Kolmogorov–Smirnov does not allow us to reject the hypothesis null at a 5% level of confidence. Second, the growth rate values of wild-type and *pspBCG* triple mutant observed from 4 to 8 h were compared using a Mann–Whitney test. The results observed under anaerobiosis are the only condition for which we could reject the null hypothesis (of equal growth rate) at the 5% level (P value $= 5.04 \times 10^{-06}$).

β-Galactosidase (β-gal) Assays

Activity from a single-copy chromosomal ϕ(*pspA-lacZ*) transcriptional fusion was assayed to gauge the level of *psp* expression. The ϕ(*pspA-lacZ*) transcriptional reporter fusion was introduced as a single copy into the chromosomal *att* site to retain the native *pspA* locus and permit the wild-type Psp response. Cells were grown overnight at 37 °C in LB broth containing the appropriate antibiotic and then diluted 100-fold (initial OD_{600nm} ~0.025) into the same medium (5 ml). Following incubation to OD_{600nm} ~0.4, cultures were assayed for β-gal activity as described by Miller (1992). For anaerobically grown cells, β-gal activity was measured following overnight growth at 37 °C with no shaking (initial OD_{600nm} ~0.010 and after incubation ~0.4). For all β-gal assays, mean values of six independent assays taken from technical duplicates of three independently grown cultures of each strain were used to calculate activity. The data are shown as mean values with standard deviation error bars.

Western Blot Analysis

For the western blot analysis, bacterial cells were collected (1 ml) at OD_{600} ~0.4. Proteins were separated on 12.5% sodium dodecyl sulfate–polyacrylamide gel electrophoresis and transferred onto polyvinyldene fluoride membranes using a semidry transblot system (Bio-Rad). Western blotting was performed as described (Elderkin et al. 2002) using antibodies to PspA (1:1,000 dilution with anti-rabbit secondary antibodies) and antibodies to pIV (a gift from M. Russel) (1:10,000 with anti-rabbit). The proteins were detected using ECL plus Western Blotting Detection Kit according to manufacturers' guidelines (GE Healthcare). Images were captured in a FujiFilm—intelligent Dark Box by an image analyzer with a charge-coupled device camera (LAS-3000).

Results

Maintenance of *psp* Genes

In this study, we investigate the properties of the Psp system across all the bacterial species present in the NCBI database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). We first determine the orthologs for the Psp system proteins (see Materials and Methods). The number of genomes containing an ortholog varies considerably between proteins, with a 4-fold difference between the maximum and minimum (fig. 2*a*, white bar). The most abundant ortholog sets are found, in decreasing order, for *pspE*, *pspC*, and *arcB*, whereas *pspF*, *pspG*, and *pspD* are the least frequently shared genes.

These results contain information from species that have only a single identifiable ortholog to one of the genes associated with the *E. coli* Psp system. Because our interest focuses on the system level and how the Psp stress response system has evolved over time (including through gain and loss of genes), we therefore restrict most of the remaining analyses to organisms containing at least two orthologs to the PspA–G proteins; ArcA and ArcB are known to be involved in several processes (Kim and Cho 2006; Malpica et al. 2006) and are therefore treated differently here, as we expect their involvement in several pathways to have also affected their evolutionary history. This filtering of species impacts differently on each ortholog set (fig. 2*a*—black bar). This does not affect the results for proteins PspD, F, and G as these are rarely found alone, whereas the number of species containing PspE and/or PspC drops quite considerably (by almost 50% for PspE), as these genes often

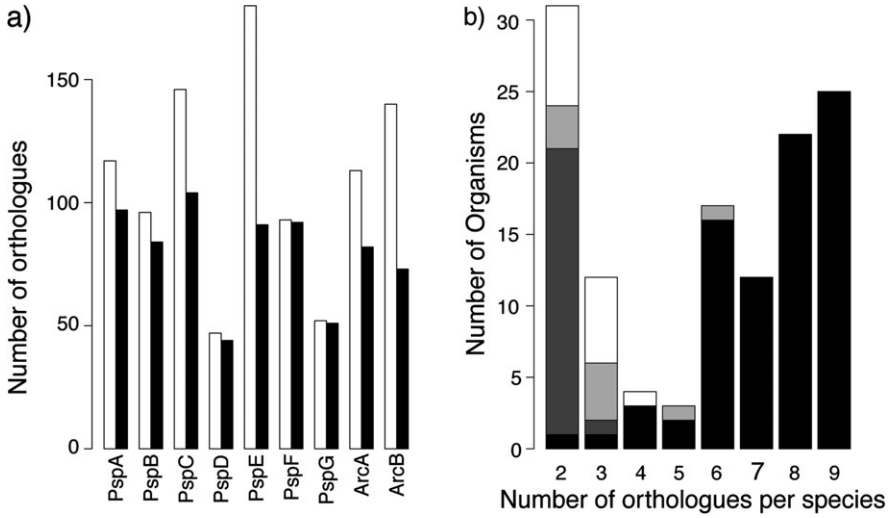

Fig. 2. Distribution of orthologs of proteins in the Psp system orthologs search results. (*a*) Number of orthologs identified for each protein of the Psp system identified in the complete species set (white bar) and in the species of interest containing at least two orthologs of the PspA to G set (black bar). (*b*) Number of orthologs per species for the species of interest. The colors correspond to the bacterial clades (black: gamma-protobacteria, dark gray: Firmicutes, light gray: alpha-protobacteria, and white: others).

occur on their own. This observation already suggests that in many species, cell membrane stress is dealt with by a very different system; it furthermore indicates that some orthologs of PspE and PspC can fulfill functions unrelated to their respective full roles in *E. coli* and other enterobacteria. The unorthodox TCS system ArcB/ArcA is maintained widely across this phylogenetic panel, underlining its important role in other stress response processes. Interestingly, however, detailed analysis shows that even TCSs can break apart over evolutionary time scales; for example, we find genomes where *arcB* but not *arcA* (and vice versa) orthologs can be identified.

In figure 2, we illustrate the distribution of the system's constituent genes across all species (with at least one and two orthologs, respectively). The distribution of the numbers of orthologs present in the 127 species takes the shape of an asymmetrical "U," where the most frequently observed results are for species to have either two or all nine genes in their genomes; only a few species have four or five orthologs of the *psp* genes. Over 70% of species with only two orthologs are "Firmicutes" (very few Firmicutes have more than two orthologs). For the "gamma-proteobacteria," on the other hand (which include the *Enterobacteria* such as *E. coli*, as well as several other biomedically relevant pathogens, e.g., *Salmonella*, *Yersinia*, and *Vibrio*), we find that almost the complete system is present in all the species (fig. 2*b*). A single "alpha-proteobacterium" is the only other species with six or more orthologs. The "U" shape observed in figure 2*b* is thus entirely due to the unequal distribution of genes among the different bacterial clades; for each individual clade (e.g., Firmicutes or alpha-proteobacteria), the corresponding distribution tends to be simpler; system composition (or mere presence) is therefore closely associated with the clade structure of the bacterial species considered here.

Moreover, and from a functional perspective probably more interestingly, we find that the manner in which orthologs are shared is not random (see supplementary fig. S1, Supplementary Material online): for example, approximately 65% of species with only two orthologs contain *pspC* and *pspE*. Conversely, a large proportion of species (15 of 22; P value $< 10^{-8}$) containing eight orthologs lack only *pspE*.

Correlated Evolution and Loss and Gain of Genes in the Psp System

We have tried to identify correlated evolution in the patterns of presence and absence for all pairs of genes in the Psp system. This would, for example, hint at the presence of one gene being necessary for another gene to fulfill its function (at least as far as relating to within the Psp response system) and hence to be maintained in the genome by natural selection (Hershberg et al. 2007).

We used the BayesTraits package to identify coevolving pairs of genes (Pagel 1994; Organ et al. 2009) (see Material and Methods). We find statistically significant evidence against the independent model only for the pair *pspA–pspF* (P value < 0.05; after correcting for multiple testing using FDR). Thus, all other gene pairings (interestingly also the two-component system *arcB–arcA*) show no statistical evidence for having correlated patterns of presence/absence. This result does, however, hint that the *pspA–pspF* subsystem may define a functional core of the phage shock response, consistent with the minimal Psp system found to be sufficient for stress response (Jovanovic et al. 2009).

Maintenance of the *psp* Operon

Crucially, we find that the level of gene sharing across the species, including the somewhat special position held by *pspE*, appears to be linked to, or reflected by, the emergence of the *psp* operon (see fig. 1). This is confirmed by jointly considering together the phylogeny, genomic organization, and putative network architecture as illustrated in figure 3 (see supplementary fig. S3, Supplementary

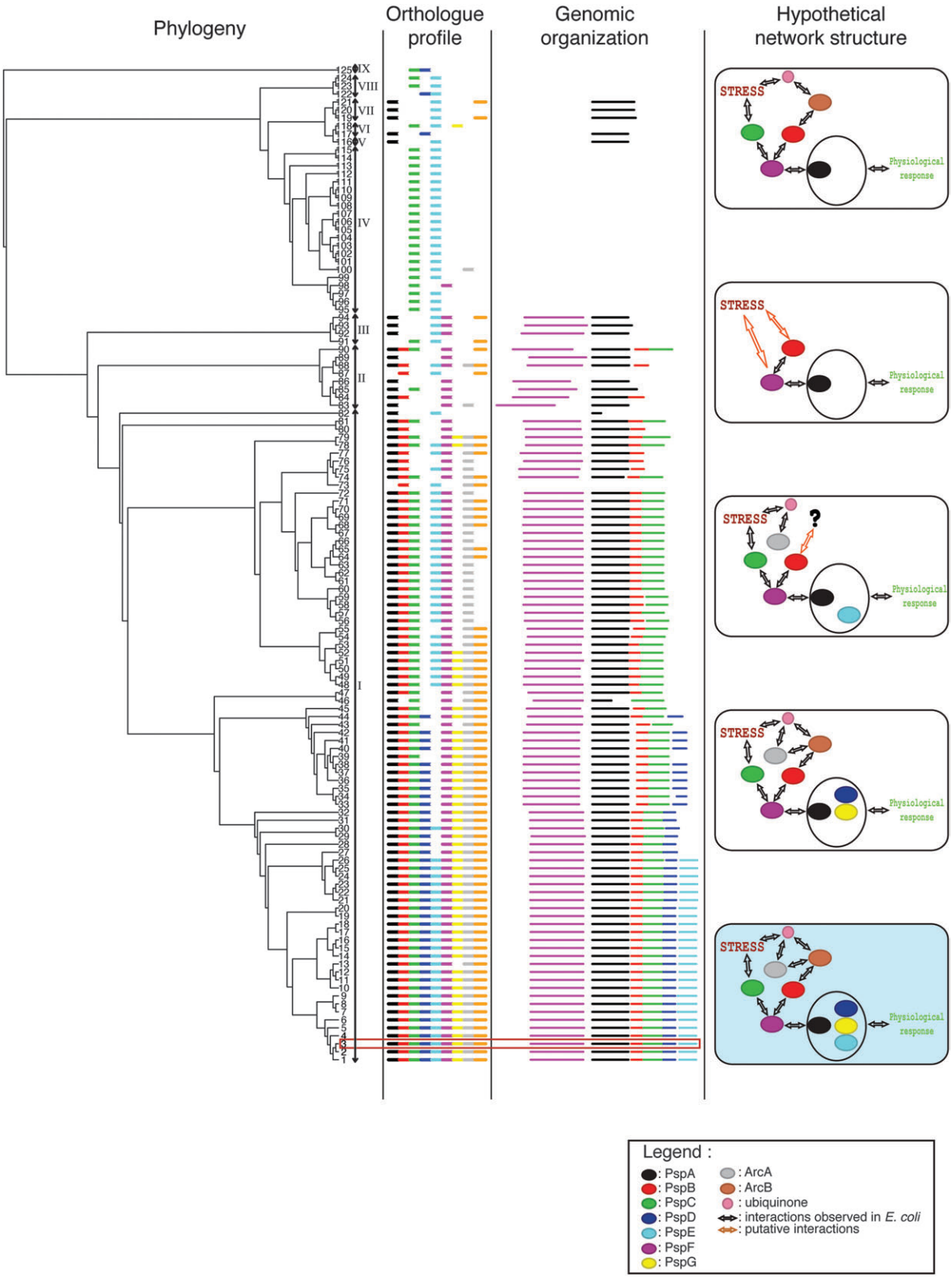

Fig. 3. Profile of presence/absence and configuration of the Psp system orthologs. Representation of the Psp system orthologs obtained for the species of interest. The orthologs for each species are organized based on their position in the phylogenetic tree. From left to right: 1) Phylogenetic tree inferred from the 23S rRNA sequences, 2) presence/absence profile of the nine proteins of the Psp system from PspA to ArcB, 3) genomic organization of *pspA* to *pspF* orthologs according to the position of the first *pspA* gene nucleotide, and finally, 4) putative inferred network organization of the Psp system for different observed set of orthologs. The color code is identical to the one used in figure 1. The correspondence between the species name and number can be found in the supplementary table S1 (Supplementary Material online).

Material online). PspE appears to be shared more widely than many of the other Psp proteins, yet it exists in two different groups: One is highly conserved compared with the *E. coli* PspE sequence, whereas the other group is made up of a set of sequences that are considerably more divergent from their *E. coli* ortholog than the other *psp* gene orthologs from the same species. Sequence similarity with respect to the *E. coli* version does not decay progressively but presents a sharp drop in observed identity (see also Extended *psp* Operon and Gene Linkage section). This dichotomy correlates perfectly with the presence of *pspE* inside the *psp* operon and is not observed for the other proteins involved in the Psp response. For other proteins, we do observe a more gradual and slower decay in levels of observed sequence similarity and no such strong dependence on the genomic (operon) organization. This suggests that PspE's role in the stress response is tightly coupled to being part of the operon. It is easy to imagine that recruitment of a gene into an existing operon, which through coordinated expression (directed by the use of common transcription factors), could result in a gain (or redirection) of function compared with the original gene.

As expected, the absence or presence of genes alone gives us only a partial picture of the evolutionary history of biological systems. But this analysis does suggest the synergistic effects of having complete systems (compared with *E. coli*): once a set of proteins working together in concert has been assembled in order to fulfill a function that affects the (Darwinian) fitness of an organism, then jettisoning one or a few of the genes from the genome may prove too detrimental for the organism.

As a result, we may expect to see this need to maintain a complete set of genes to be reflected at the operon level. In *E. coli*, the *psp* operon is 1 of 44 operons containing five genes (Pertea et al. 2009). The distribution of the number of orthologs found in other species is, in fact, fairly flat for the *psp* operon compared with the other 43 distributions (see supplementary fig. S2, Supplementary Material online).

Gain and loss of genes occur predominantly at the end of the operon; *pspA*, *pspB*, and *pspC* which are engaged in sensing stress, transducing the signal, and as an effector in the case of PspA (Model et al. 1997; Darwin 2005) (for *Yersinia enterocolitica*, PspB and PspC are also effectors; Maxson and Darwin 2006b) are generally kept together in the species considered here. The less well functionally characterized genes (knock outs of *pspD*, *pspE*, as well as *pspG*, which resides outside the *psp* operon, exhibit no growth phenotypes and can maintain the pmf during stress) are absent from other species more often than *pspABC* and *pspF*. PspD may have an effector function that overlaps with that of PspA (Jovanovic et al. 2006). The PspA and PspD effector functions are distinct from the functions of PspG or PspE, which appear to be involved in modulating cellular metabolism under stress conditions (Adams et al. 2002; Jovanovic et al. 2006).

The observed pattern suggests that the Psp response system has become more nuanced and complex over time in *E. coli*. In particular, *pspE* appears to be a recent addition to the Psp response machinery, which in other bacteria fulfills a different role.

Extended *psp* Operon and Gene Linkage

The analysis of the *psp* operon structure suggests that *pspA–E* may have not been the only genes making up the *psp* operon. Many approaches have been developed to predict operons; we used one of the most commonly used and simplest measures to identify potential extended versions of the *psp* operon (Brouwer et al. 2008) (see Material and Methods). Our objective here is just to identify genes potentially associated with the *psp* operon genes in the sequenced species and not explicitly predict operons ab initio.

This analysis reveals three interesting aspects of the *psp* operon's evolutionary history (see supplementary fig. S6, Supplementary Material online). First, all the genes that belong to the *psp* operon in *E. coli* are potentially also part of other operon structures in at least one of the studied species. Second, several species contain *psp* operon structures with intergene fragments within the operon that are longer than 100 bp. Finally, some of the *psp* operon structures formed by *pspAB*, *pspABC*, or *pspABCD* are potentially extended by additional genes. In total, 32 of the 81 *psp* operon structures identified in this study contain at least one additional gene, *ycjX*. It had already been reported that *ycjX* and *ycjF* could be part of the *psp* operon structure in a number of bacterial species; however, no systematic analysis of the association between these two genes and *psp* operons had been performed up to now (Darwin 2005).

Orthologs for *ycjX* and *ycjF* are identified in almost all species also containing orthologs for at least *pspA* and *pspB*. In almost all the species studied, these two genes are adjacent. However, the physical organization of *ycjX* and *ycjF* in relation to the *psp* operon presents a large level of variability, especially in comparison with the *psp* operon genes or the *ycjXF* pair itself. For example, the species presenting an ortholog for *pspA*, *pspB*, and *pspC* but none for *pspD* present cases with *ycjX* and *ycjF* next to *pspC* with an intergenic distance shorter than 100 bp, others with an intergenic distance larger than 100 bp but *ycjX* still next to *pspC*, and finally cases with *ycjX* and *xcjF* in different chromosomal locations. For the species presenting orthologs for *pspABCD* and *pspABCDE*, the same diverse set of profiles can be observed for these additional genes.

Coordination with *psp* Genes Outside the *psp* Operon

A potential clue as to how *pspG* was recruited into the Psp response system is provided by *Aeromonas salmonicida A449*, where the *pspG* ortholog is found immediately adjacent to a sequence stretch containing paralogs to *pspA* and *pspC*; here, the distances between the coding sequences for the *pspA* and *pspC* paralogs and the *pspC* paralog and the *pspG* ortholog are 13 bp and 47 bp, respectively. This proximity together with experimental observations that the three genes are in *E. coli* coordinately regulated by *pspF* indicates that they could form their own operon. The orthologs for *pspA* and *pspC* that are detected using

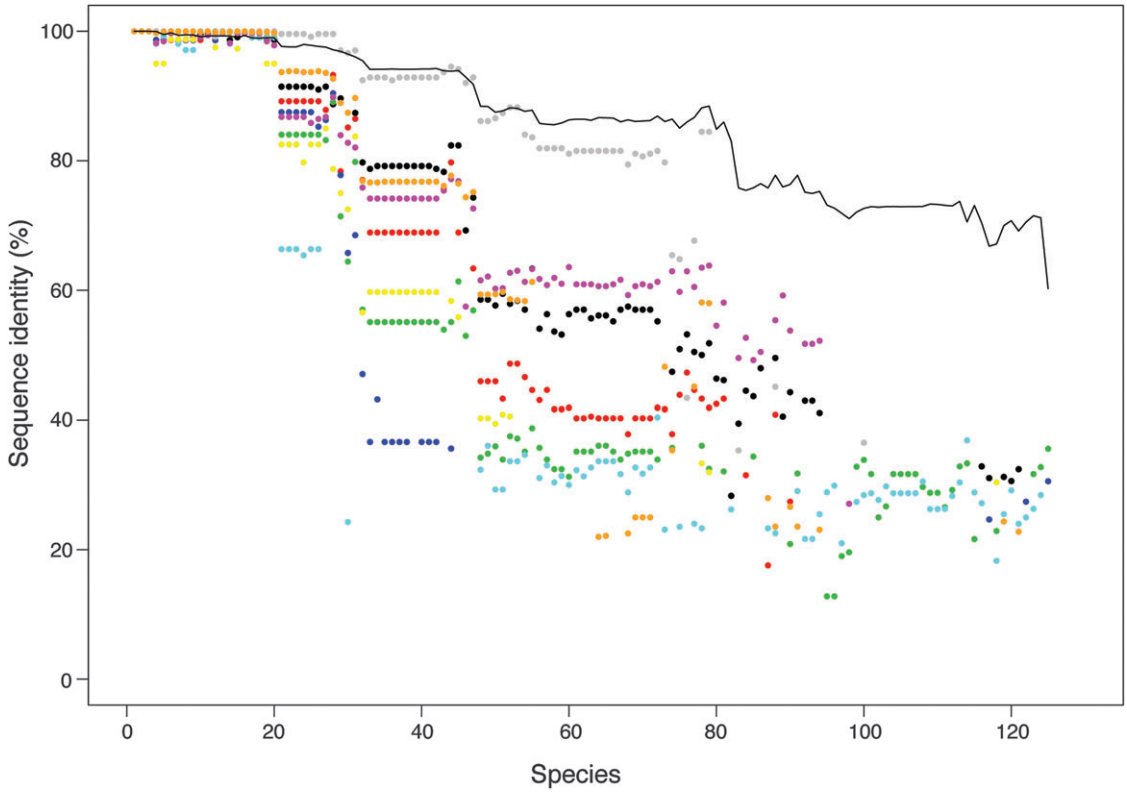

Fig. 4. Similarity variation of orthologous proteins in the Psp system. Each colored dot represents the sequence identity (in percent) of the identified ortholog against the corresponding *Escherichia coli* K12 protein sequences. The position of a species in the *x* axis corresponds to the position of the species in the 23S rRNA phylogenetic tree as presented in figure 3. The black line corresponds to the identity percent of the 23S rRNA sequences of the different species against the *E. coli* K12 genomic sequence (see Materials and Methods). The color code is identical to the one used in figure 1. The correspondence between the species name and number can be found in the supplementary table S1 (Supplementary Material online).

best reciprocal Blast occur elsewhere in the genome of *A. salmonicida A449* and are to either side of a clear *pspB* ortholog (the published annotation of the genome, however, fails to identify the correct *pspA* and *pspC* orthologs). This establishes an extant transcriptional link between *pspG* and *pspAC*, which opens up the intriguing possibility that *pspG* acquired its association with, and role in, the Psp response through historical association with an operon containing paralogs to *pspA* and *pspC*.

Sequence Similarity and dN/dS

Until now we have treated the orthologs, with the exception of PspE, as identical objects, which they are clearly not. The percentage sequence identities of the different orthologs were computed with respect to *E. coli* and plotted based on the position of the species in the phylogenetic tree (phylogenetic relationships are based on the 23S rRNA sequences extracted for all the species) (fig. 4). As expected, the relative sequence identity is correlated with the evolutionary distance to *E. coli*. In any given species, however, relative divergence to *E. coli* can differ quite considerably between the different Psp and Arc orthologs. ArcA is the most conserved protein with, except for 4 cases of 82, a minimum of 60% amino acid sequence identity and is followed closely by the two pivotal *psp* genes, PspA and PspF. The proteins with the largest number of orthologs (PspC and PspE) have the lowest levels of identity across the panel of sequenced bacterial species. In PspE, this is probably related to the potential alteration in its functionality once the gene is recruited into the psp operon.

Using standard evolutionary arguments, we can infer whether or not the evolution of a protein is in agreement with the behavior expected from neutrally evolving genes, by considering the ratio of nonsynonymous to synonymous substitution rates for all orthologs of the corresponding *E. coli* genes (see fig. 5) (Yang and Bielawski 2000). The dN/dS ratios are mostly smaller than 1, which suggests the action of purifying selection. This is to be expected from genes, once acquired, which fulfill important functions for the organism and thus need to be maintained. Only for PspF do we observe large values of this ratio, which is indicative of adaptive evolution in a few species.

The link between sequence similarity and the dN/dS measure is illustrated further for PspF and PspE in figure 6 (the corresponding plots for the other protein sequences are shown in supplementary fig. S5, Supplementary Material online) where we show these two quantities for all pairwise species comparisons. For a large proportion of the gene pairs, the sequences are too divergent to allow the reliable calculation of a relevant dN/dS ratio, and the majority of pairs with a computable ratio involve closely related species (which concentrate around the diagonals

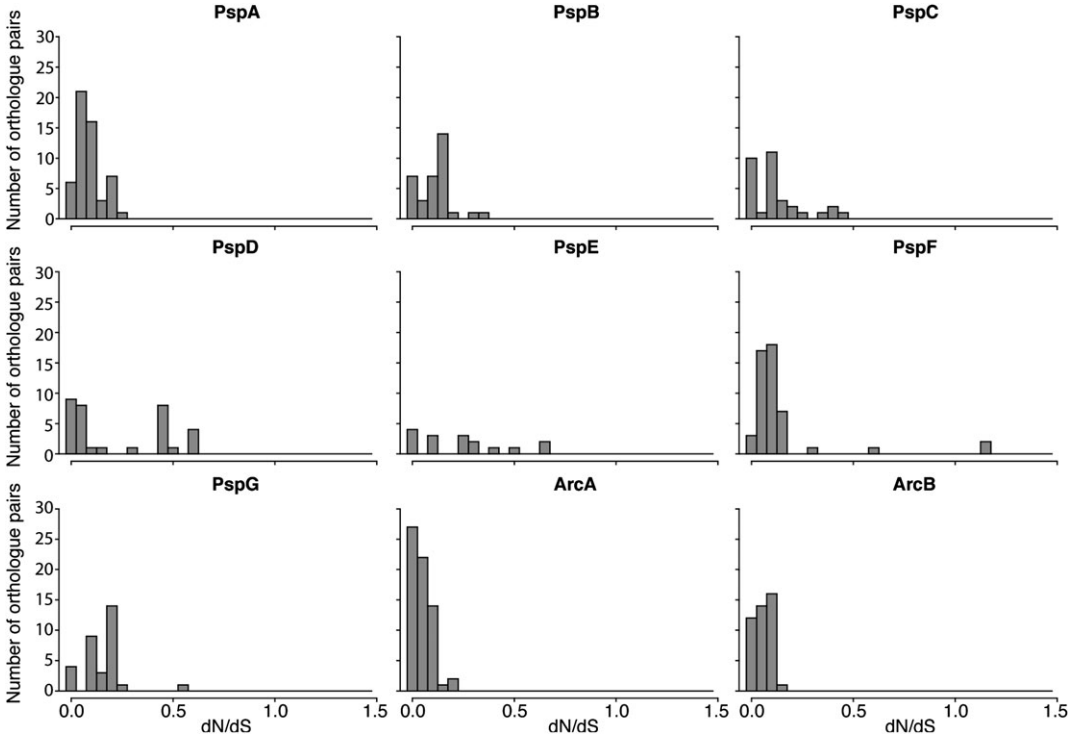

Fig. 5. Estimation of the selection pressure according to *Escherichia coli* K12. Histograms representing the dN/dS ratio estimates between the *E. coli* K12 sequences and all identified orthologs (see Materials and Methods).

in fig. 6). But pairs of species with high levels of sequence identity tend to show strong evidence for purifying selection and vice versa. The block structure in figure 6 clearly reflects the evolutionary relationships already apparent in the phylogeny in figure 3: at the sequence level, the clade structure is also clearly reflected.

Intergenic Regions

There is increasing evidence that the effects of selection extend well beyond the coding sequences (Hurst 2009; Paget and Helmann 2003)—as had previously been posited. The controlled and coordinated expression of genes is largely dependent on features in the noncoding sequence, predominantly the promoter and termination sites. We focus here on the regions separating *pspF* from the *psp* operon and the intergenic regions within the operon (supplementary fig. S4, Supplementary Material online).

The region between *pspF* and *pspA* contains the regulatory sequences that in *E. coli* control the repression of *pspF* expression and the activation of the genes in the *psp* operon (Weiner et al. 1991; Weiner et al. 1995; Jovanovic et al. 1996; 1997, 1999; Dworkin et al. 1997; Jovanovic and Model 1997). In *E. coli*, this region contains four annotated promoters; according to present experimental evidence (Jovanovic et al. 1996), three of these correspond to *pspF* and one to *pspA*. As expected for a promoter region, the overall sequences are highly divergent overall, but the sets of closely related species display similar sequence profiles. In the overall alignment, none of the three *pspF* promoters show a conserved profile beyond the enterobacteria. The multiple alignment presents a different result for the *pspA*/*psp* operon promoter. The signatures of the σ^{54}-dependent promoter (Wigneshweraraj et al. 2008), −24 and part of the −12 box, are practically identical in all the species considered here. A detailed analysis of the −12 box reveals that the classic GC motif is replaced in almost all species by the unusual GT motif. All those results are congruent with the expected role of PspF as a transcription factor of *pspA* and the other genes of the *psp* operon in these species.

The above is, however, not the only regulatory sequence architecture found among the panel of species; in *Yersinia*, the upstream region of the *psp* operon, for example, contains an additional σ^{70}-dependent promoter, which adds additional complexity and flexibility to the transcription regulation in this system (Maxson and Darwin 2006a). However, in *E. coli* and *Salmonella*, the regulatory control of the operon (with the notable exception of *pspE* discussed in more detail below) is known to be entirely σ^{54} dependent (Lloyd et al. 2004), as is also clear from our analysis.

The control of *pspE* expression appears to be subtly different from the rest of the genes in the *psp* operon. In *E. coli* and *Salmonella typhimurium*, *pspE* has its own σ^{70}-dependent promoter site, which allows *pspE* to be expressed independently from the other genes. However, when under control of the operon σ^{54} promoter, *pspE* is transcribed as part of the polycistronic *psp* operon mRNA. Importantly, the stress-induced transcription (under σ^{54} control) of *pspE* will be markedly increased (5- to 10-fold) compared with the basal expression levels observed under normal growth conditions, corresponding to combined σ^{70}/σ^{54}-dependent transcription.

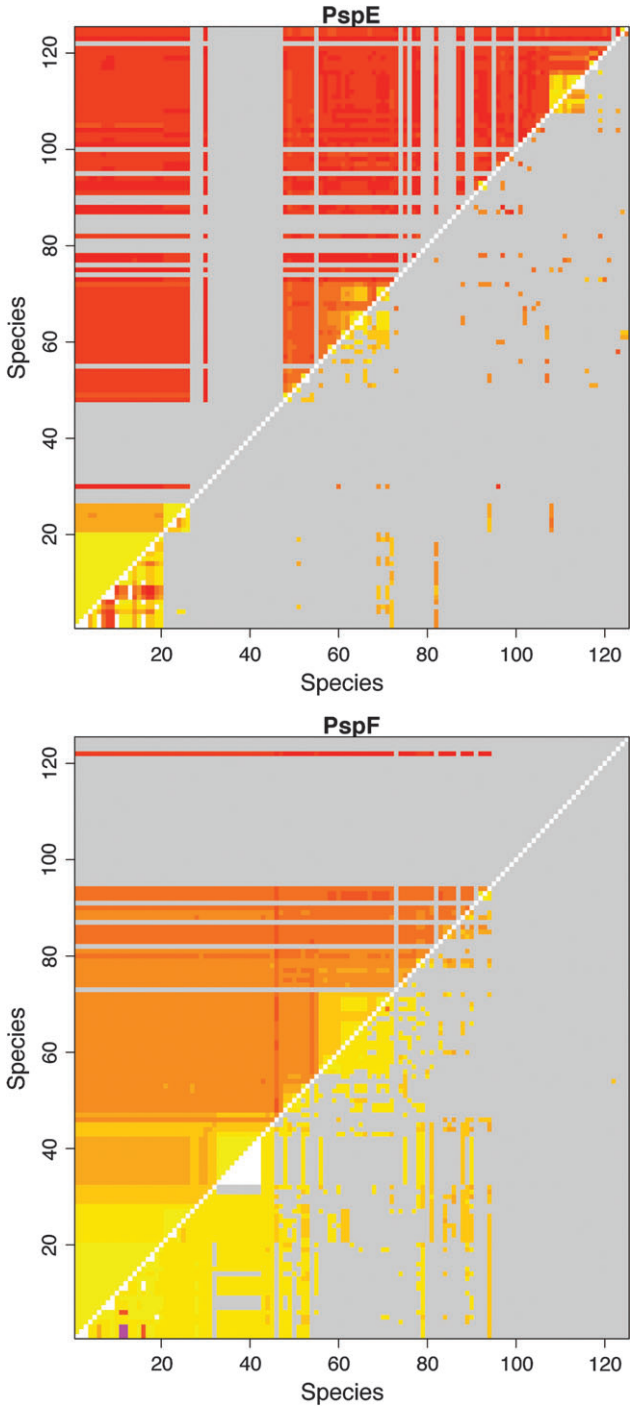

Fig. 6. Estimation of the selection pressure for all ortholog pairs for PspE and PspF. These heat-map matrices represent two types of information: the results of the dN/dS ratio estimation (lower triangle) and the protein sequence identity (upper triangle) computed for each pair of sequences from the PspE (top) and PspF (bottom) ortholog sets. The positions of the species in the matrix correspond to their positions in the unrooted 23S rRNA phylogeny. The values of the dN/dS ratio and identity percent are coded using the following color codes: 1) identity percent values range from 0% to 100% and correspond to a color gradient going from red to yellow; 2) dN/dS ratios between 0 and 1 correspond to a color gradient going from yellow to red, whereas values larger than 1 are in purple (the ratios can take any values from 0 to 10, an artificial upper limit set by the program used to do the estimation). If the sequences are too close to allow reliable estimation of the dN/dS ratio, the color is white. Finally, gray fields correspond to positions in the matrix for which no orthologs are present in the respective species or where the pair of orthologs is too divergent to allow for reliable estimates of the dN/dS ratio.

The predominating features in the intergenic regions inside the psp operon are 1) the transcription termination site between *pspA* and *pspB* and 2) the boundary between *pspB* and *pspC*. The former results in stochastic dissociation between the transcriptional machinery and the chromosome, which results in markedly reduced transcription of *pspB–E* compared with *pspA* (at a ratio of roughly 1:4). Potentially PspBC acts on a PspA–F complex containing six subunits of PspA, offering some rationale for the differences in expression levels (Joly et al. 2009). The latter, however, appears to reflect a somewhat more subtle relationship between the two genes than is apparent from their respective amino acid sequences. In many species, the *pspB* and *pspC* sequences partially overlap: In *E. coli*, the stop and start codons of *pspB* and *pspC* overlap; in other species (such as *Shewanellas ANA-3* or *Shewanella baltica OS 185*), the actual coding regions overlap up to approximately 10 nt. The resulting cotranslation will quite generally be helpful for protein complex formation, which in turn enables pairs or groups of proteins to fulfill any functions that depend on their joint action. Detailed analyses of all *pspB–pspC* junction regions furthermore reveals that in *Sodalis glossinidius str. "morsitans,"* a single gene is present in the region normally occupied by *pspB* and *pspC* (incorrectly annotated as an ortholog to *pspB* in the database); further inspection reveals that the predicted corresponding protein product is a chimera consisting of two domains, one belonging to PspB and one belonging to PspC.

Psp Response of the Reduced Psp Regulon

To address the question of how effective the Psp response will be in *E. coli* cells with a significantly reduced Psp regulon, we constructed a strain lacking *pspBC* and *pspG* genes (the probable first genes added to an hypothetical ancestral Psp system formed by *pspA* and *pspF*) and determined levels of *psp* induction and expression of PspA in the absence or presence of secretin pIV, the inducer of cell wall stress and under different oxygen-related environments (aerobic, microaerobic, and anaerobic growth conditions).

Importantly, the *psp* induction and expression of PspA in *pspBCG* triple mutant is markedly reduced under aerobiosis and microaerobiosis (fig. 7*a* and *b*). Because the same has been shown for induction of *pspA* in a *pspBC* mutant (Jovanovic et al. 2009), these new results support the view that the PspG protein does not contribute to the regulation of *psp* induction (Lloyd et al. 2004). Under anaerobiosis, the expression of PspA shows no difference between the wild-type and the triple mutant. Taken together, these results and the observation that pIV production in *ΔpspF* cells (occurring in the absence of Psp proteins) significantly impairs the cell growth (Jovanovic et al. 2006) suggest that expression of PspA (and perhaps PspD and PspE) is sufficient to maintain robust cell growth under *psp*-inducing

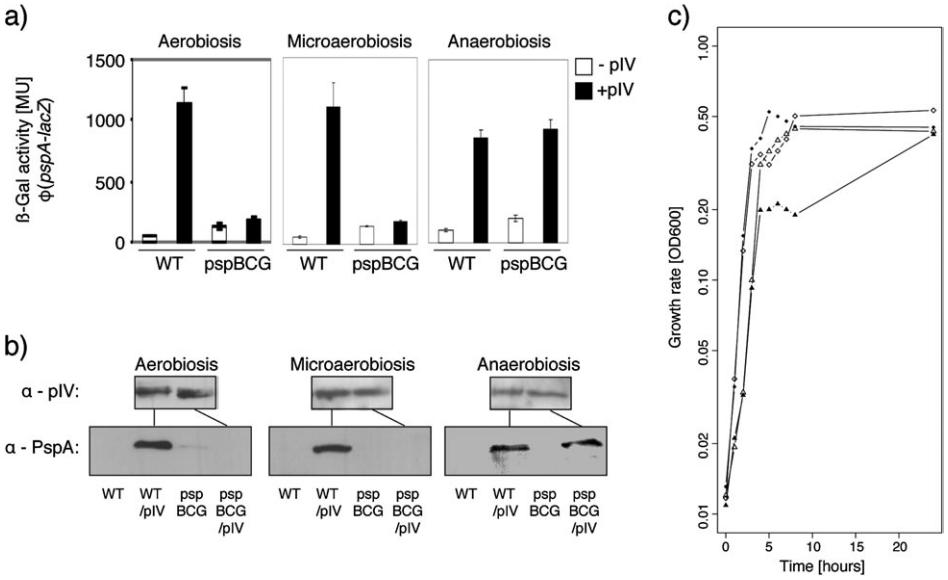

Fig. 7. Psp response of a *Δ*pspBC *Δ*pspG mutant. (*a*) Effect of continual secretin production on the psp induction of *Δ*pspBC *Δ*pspG mutant under different growth conditions. The expression of pspA–lacZ chromosomal transcriptional fusion was measured (*β*-gal activity, see Materials and Methods) in wild-type (WT; MVA44 [Jovanovic et al. 2006]) and *Δ*pspBC *Δ*pspG (pspBCG; MVA103) strains in the absence (−pIV) or presence (+pIV) of pIV secretin under aerobic, microaerobic, or anaerobic growth conditions. (*b*) Expression of PspA in a *Δ*pspBC *Δ*pspG mutant producing pIV secretin under different growth conditions. The samples from all strains used in (*a*) were collected at OD_{600} ~0.4, and expression of PspA protein was determined using western blot and antibodies against PspA (*α*-PspA). The expression of pIV in the corresponding samples was detected using western blot and antibodies against pIV (*α*-pIV). (*c*) Effect of continual secretin production on the growth of *ΔpspBC ΔpspG* mutant strain under anaerobic growth conditions. The plasmid pGJ4 with leaky expression of pIV secretin was transferred into wild-type (WT; MG1655) and *ΔpspBC ΔpspG* (pspBCG; MVA49) strains (genotypes indicated on the right-hand side). The strains with or without pGJ4 were grown under anaerobic growth conditions as described in the Materials and Methods and optical density was measured at hourly intervals. The data are from a single experiment in which all strains were tested simultaneously (the experiment was carried out on three separate occasions to ensure reproducibility—see supplementary fig. S7, Supplementary Material online).

stress conditions (pIV production) at a level comparable with wild-type cells with the intact Psp regulon.

Based on these observations, we looked for the influence of the presence or absence of *pspB*, *C*, and *G* on cell growth. The results show that production of pIV reduces growth to similar extent in the wild-type and the *pspBCG* triple mutant cells under aerobic and microaerobic growth conditions compared with nonstressed cells (supplementary fig. S7, Supplementary Material online). This is in line with the observation that pIV production in cells lacking *pspBC* does not impair the growth in comparison with stressed wild-type cells (Jovanovic et al. 2006) and establishes that *pspG* does not contribute to the observed growth rates. Notably, in anaerobiosis, the *pspBCG* triple mutant cells show better growth (and so have a potential fitness advantage over wild-type cells) after 5-h growth under stress conditions (fig. 7*c*), confirming that the induction of the *psp* genes upon production of pIV does not depend on the presence of PspBC under anaerobic growth conditions and demonstrating a growth condition–dependent advantage of having the full complement of *psp* genes (Jovanovic et al. 2009). The statistical tests applied are described in the Bacterial Strains and Growth Conditions section of the Material and Methods.

PspG may be important for fine-tuning the metabolism toward anaerobic respiration under stress in aerobiosis and microaerobiosis (Jovanovic et al. 2006), indicating a growth condition–specific role for PspG in addition to the demonstrated conditional use of PspBC. We propose a potential increase of selective pressure upon wild-type cells compared with *pspBCG* triple mutant under stress growth conditions in anaerobiosis reflecting an advantage in aerobic growth of having PspBC and PspG. It should be noted that the results obtained with the *pspBCG* triple mutant are related to the *E. coli* version of the remaining *psp* genes, mainly *pspA* and *pspF*. They could be different from their ancestral version. This could result in a difference of phenotype between an organism containing only *pspA* and *pspF* as they were in ancestral species (not containing *pspBCDE* and *G*) or as they are in *E. coli*.

Discussion

When considering the evolutionary process that has given rise to the complex molecular signaling machinery of, for example, stress response mechanisms, it is important to consider the need of bacteria (and other organisms) to interact with the changing environment. This, of course, exerts selection pressure on the components making up the system (Pinney et al. 2007; Ratmann et al. 2009). In addition to selection, other factors such as genetic drift also affect the evolution of biological systems (Lynch 2007a, 2007b). It has been argued that nonadaptive processes can give rise to systems that exhibit comparable levels of complexity to those observed in biological organisms (Lynch 2007a). In

the absence of population-level data, we must therefore be careful in interpreting the results of evolutionary analyses.

Patterns of coinheritance suggest strongly that the constituent genes do not evolve independently when considered as a functional system (here, details of the system's function do not matter in the first instance): if we think of the proteins as useful tools, then an organism may just as well jettison the whole toolbox once too many individual tools have been lost. There are, however, subtle differences that require a more nuanced description. These are presumably due to the intricacies of the molecular machinery underlying the Psp response, which are determined by the multiple roles played by several genes: first of all, the TCS consisting of ArcA and ArcB is also involved in different processes (Malpica et al. 2006). Second, PspA plays a pivotal role in receiving the signal and releasing the negative control upon PspF once the stress has been sensed and transduced by a combination of the PspB and PspC sensors (reviewed in Model et al. 1997; Darwin 2005); furthermore, PspA acts as an effector involved in repairing cell wall damage and contributes to reestablishing pmf (Kleerebezem et al. 1996; Jovanovic et al. 2006; Kobayashi et al. 2007; Engl et al. 2009;). PspB has an additional role in mediating quantitative effects of the Psp response through interactions with ArcB, integrating the ArcAB two-component system into the Psp response machinery (Jovanovic et al. 2009). In fact, a mechanistic model describing only the effects of pspA, pspB/C (jointly, rather than independently), and *pspF* may suffice to capture the qualitative response.

Interestingly, the level of correlated evolution does not appear to matter when considering only pairs of genes: in such an analysis only PspA and PspF show statistical evidence for correlated evolution. However, when considering the whole system, then system-level correlated evolution—that is, beyond the level of pairs of genes—becomes apparent.

At the sequence level, there is clear evidence that all the genes have been under purifying selection; this evidence becomes overwhelming when assessed not across the whole phylogeny but on a clade-by-clade basis. In fact, using the wealth of the available information in one fell swoop may easily mask these signatures at the sequence level; here, evolutionary distance (or differences in ecological niches occupied by different species) could play the role of a confounding factor or hidden variable (Thorne and Stumpf 2007; Chuang et al. 2009).

Our results highlight the crucial role, which the operon organization has played in the evolutionary history of the Psp response system. A range of plausible mechanisms have been put forward to explain operon organization and evolution: 1) operons may form to ensure that proteins with related function or involved in the same process are maintained as a coherent unit; 2) being under shared transcriptional control maintains balanced mRNA expression levels and stochiometries; 3) the order in which genes are transcribed and polycistronic mRNA is translated could ensure that proteins needed first or upstream of the others are made available in the right order; and finally, 4) the (somewhat controversial) "selfish operon" model assumes that the co-localization/coexpression may not be an advantage to the organism containing the operon but rather directly to the operon genes (Fondi et al. 2009).

All of our results clearly support the notion that the operon grew in the head-to-tail direction from the 3′ end of *pspA* from the *pspF pspA* divergent transcription unit. Importantly, the physical organization of the genes in the operon is not simply a function of where they occur in the pathway/network (see fig. 1*c*). Rather, in light of the complete set of results discussed above, the most likely scenario is that the system has evolved around a core cell wall–stress response consisting of only PspA and PspF, which by themselves are known to be able to induce a functional stress response under specific conditions (Jovanovic et al. 2009). These genes form a minimal system that can elicit a response where PspA is the sensor, regulator, and effector and PspF the regulated gene activator. Addition of PspB and PspC (or perhaps a chimeric protein combining domains of both) would then have fine-tuned the initiation and modulation of the response, which agrees well with what is known about their function in *E. coli* (Weiner et al. 1991; Gueguen et al. 2009) and the experimental results presented above. Such a configuration retains the essential control function at the transcriptional level for PspF. Finally, PspD and PspE must then have been added, leading to the operon structure now observed in *E. coli K12* and other species.

The search for genes that are potentially part of the *psp* operon but not observed in *E. coli* revealed that few species might contain extra genes within their respective *psp* operons. In all species studied and presenting an alternative version of the *psp* operon, the same set of genes were observed (*ycjX* and *ycjF*). However, the genomic organization reveals that these extra genes show a much weaker level of physical association to the *psp* operon than do the *psp* genes. Although we cannot exclude the possibility of a functional link between *ycjXF* and the Psp system, we found that almost each sub *psp* operon (with the exception of *pspAB*) could be observed without *ycjXF* in close vicinity. The phylogenetic profile of genomic organization suggests that events of physical association or dissociation occurred independently.

According to RegulonDB (Gama-Castro et al. 2008), *ycjX* and *ycjF* are part of a separate transcriptional unit, verified experimentally in *E. coli*, composed of three genes (*ycjX*, *ycjF*, and *tyrR*). The profile of ortholog presence/absence, within the 127 species studied here in detail, reveals that these three genes have almost identical profiles. At a genomic level, almost all *ycjX* are next to *ycjF* and a large proportion of them are next to *tyrR*. *ycjX* is reported to be associated to a σ^{32} promoter. This promoter is annotated as a regulator in the heat-shock response system (Nonaka et al. 2006). The Psp system, although regulated by a different promoter, is also activated during heat-shock stress. This represents a potential functional link between the two systems and might explain the observed association between them. However, the Psp system is also

known, at least in *E. coli*, to respond to a large variety of other stresses providing situations during which one system will be needed and not the other, thereby reducing the potential need for operon-like expression regulation and structure. Interestingly, early analyses of *psp* transcription suggest that product of the σ^{32}-controlled heat-shock system could act to suppress *psp* expression (Weiner et al. 1991). These results indicate that in *E. coli*, the *psp* operon and *ycjX* and *F* are expressed under different conditions and thus are perhaps less likely to interact functionally. A detailed analysis of the promoter region of the different *ycjX* and *F* orthologs could help us to understand the potential relationship between these genes and the Psp system.

Our present functional knowledge of PspD, PspE, and PspG is insufficient at the moment, but we conjecture that their role may be primarily as effectors because a contribution to regulation can be ruled out (Model et al. 1997; Jones et al. 2003; Lloyd et al. 2004). A subtle effect of PspE may already be expected from the observation that it is shared widely across the bacterial species considered here, especially in species lacking any of the other *psp* genes. The present evidence implies that its role in the Psp response occurred concomitantly with its recruitment into the *psp* operon. Interestingly, the physical linkage between an ortholog of *pspG* and a paralog of *pspA* in *A. salmonicida A449* suggests that recruitment of *pspG* into the Psp response was also initiated indirectly, perhaps through coexpression due to shared transcriptional regulation. There is also recent evidence that, in *E. coli*, PspC, besides with PspA, directly interacts physically with PspG (Jovanovic et al. 2010), supportive of a particular relationship between PspA, C, and G within the Psp system.

The ArcB/ArcA two-component system is involved in several biological signaling and stress response processes, and when and how it was recruited to contribute to the Psp response cannot be inferred from our analysis, but again, the role here is to modulate the strength of response without being essential to it (Jovanovic et al. 2006, 2009).

If functional orthologs can be identified, then we may be able, with due care, to transfer our knowledge of system-level processes between model and nonmodel organisms. But what happens if we have some species for which only a subset of relevant orthologs can be identified? This could either 1) be evidence that the system can operate without these genes through a (potentially) simpler mechanism; 2) falsely imply the absence of the orthologous genes when they are actually present in the nonmodel genomes but cannot be identified using standard methods; or 3) indicate that in these organisms other proteins have been co-opted to fulfill the missing roles, which have apparently unrelated functions in the model organism, for example, *E. coli* (Berg and Lässig 2006; Lee et al. 2008; Alexander et al. 2009; Xia et al. 2009; Lewis et al. 2010).

Building onto an existing (however rudimentary) stress response system (e.g., *pspA* and *pspF*) seems like a probable evolutionary strategy, where increasing sophistication and honing of the response by adopting additional proteins as response regulators and effectors results in an apparent increase in the complexity of the system. Such an attractive model receives additional support from the analysis of other enterobacterial Psp response systems. Although *pspABC* and *pspF* are generally maintained, other proteins seem to have been incorporated in the *psp* operon and have replaced *pspD* and *pspE* in, for example, *Yersinia enterolocitica* or *Idiomarina loihiensis* (Darwin 2005). Interestingly, our experimental analysis show that such reduced systems may in fact sometimes perform better than the full system under certain environmental conditions.

The coordinated operation of a system such as the Psp response is maintained by a variety of mechanisms: Having all the required genes in the same operon is the most obvious means of achieving this; having a defined regulon characterized by shared promoters is another—perhaps less restrictive—option to ensure well ordered responses; finally, physical protein–protein interactions can mediate flexible and spatiotemporally well-defined coherent responses. Our detailed analyses have shown the different ways in which the Psp response was assembled over evolutionary time courses and is maintained today in *E. coli* (and related species). These include: overlapping reading frames that increase or decrease coexpression of neighboring genes depending on how this affects the binding of RNA polymerase to DNA; proteins that may form chimeras or break up into separate entities, depending on the need to maintain coherence and physical linkage between protein domains; additional within-operon promoter sites that provide transcriptional flexibility, allowing for additional flexibility when new proteins are recruited into operons; equally, transcription termination (hairpin loops) sites inside the operon can fulfill the same role by adjusting the relative abundance of mRNA of different genes inside the same operon. This last mechanism can, for example, ensure that different proteins are produced in the required relative abundances without the need for further posttranslational modifications.

A final note of caution may be in place: The selection pressure experienced by bacteria (or more generally any organism) is much more diverse than anything probed under laboratory conditions. Functional analyses may therefore be presenting us with too simple a picture as far as the complexities of signaling and regulatory responses to environmental stress are concerned. In particular, studying stress response systems in isolation may be oversimplifying matters, especially if cross-talk is the rule rather than an exception, which we would hazard to guess.

Supplementary Material

Supplementary material, table S1, and figures S1–S8 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

For part of this work M.H. was supported through a Wellcome Trust VIP award. M.H., T.T., M.B., and M.P.H.S.

acknowledge financial support from the BBSRC, T.T. and MPHS from the MRC, X.S. from the Kwok foundation, G.J. and M.B. from the Wellcome Trust. J.W.P. from the Royal Society; M.P.H.S. is a Royal Society Wolfson Research Merit Award Holder.

References

Adams H, Teertstra W, Koster M, TommassenJ . 2002. PspE (phage-shock protein E) of Escherichia coli is a rhodanese. *FEBS Lett*. 518(1–3):173–176.

Agrafioti I, Swire J, Abbott J, Huntley D, Butcher S, Stumpf MPH. 2005. Comparative analysis of the Saccharomyces cerevisiae and Caenorhabditis elegans protein interaction networks. *BMC Evol Biol*. 5(1):23.

Alexander RP, Kim PM, Emonet T, Gerstein MB. 2009. Understanding modularity in molecular networks requires dynamics. *Sci Signal*. 2(81):pe44.

Babu M, Musso G, Díaz-Mejía J, Butland G, Greenblatt J, Emili A. 2009. Systems-level approaches for identifying and analyzing genetic interaction networks in Escherichia coli and extensions to other prokaryotes. *Mol Biosyst*. 5:1439–1455.

Berg J, Lässig M. 2006. Cross-species analysis of biological networks by Bayesian alignment. *Proc Natl Acad Sci U S A*. 103(29):10967–10972.

Brouwer RW, Kuipers OP, van Hijum SA. 2008. The relative value of operon predictions. *Brief Bioinform*. 9(5):367–375.

Chuang JS, Rivoire O, Leibler S. 2009. Simpson's paradox in a synthetic microbial system. *Science* 323(5911):272–275.

Darwin AJ. 2005. The phage-shock-protein response. *Mol Microbiol*. 57(3):621–628.

Darwin AJ. 2007. Regulation of the phage-shock-protein stress response in Yersinia enterocolitica. *Adv Exp Med Biol*. 603:167–177.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol*. 23(2):327–337.

Dworkin J, Jovanovic G, Model P. 1997. Role of upstream activation sequences and integration host factor in transcriptional activation by the constitutively active prokaryotic enhancer-binding protein PspF. *J Mol Biol*. 273(2):377–388.

Dworkin J, Jovanovic G, Model P. 2000. The PspA protein of Escherichia coli is a negative regulator of sigma(54)-dependent transcription. *J Bacteriol*. 182(2):311–319.

Elderkin S, Bordes P, Jones S, Rappas M, Buck M. 2005. Molecular determinants for PspA-mediated repression of the AAA transcriptional activator PspF. *J Bacteriol*. 187(9):3238–3248.

Elderkin S, Jones S, Schumacher J, Studholme D, Buck M. 2002. Mechanism of action of the Escherichia coli phage shock protein PspA in repression of the AAA family transcription factor PspF. *J Mol Biol*. 320(1):23–37.

Engl C, Jovanovic G, Lloyd LJ, Murray H, Spitaler M, Ying L, Errington J, Buck M. 2009. In vivo localizations of membrane stress controllers PspA and PspG in Escherichia coli. *Mol Microbiol*. 73(3):382–396.

Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO. 2009. Reconstruction of biochemical networks in microorganisms. *Nat Rev Microbiol*. 7(2):129–143.

Fondi M, Emiliani G, Fani R. 2009. Origin and evolution of operons and metabolic pathways. *Res Microbiol*. 160(7):502–512.

Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M, et al. (19 co-authors). 2008. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res*. 36(Database issue):D120–D124.

Gueguen E, Savitzky DC, Darwin AJ. 2009. Analysis of the Yersinia enterocolitica PspBC proteins defines functional domains, essential amino acids and new roles within the phage-shock-protein response. *Mol Microbiol*. 74(3):619–633.

Hershberg R, Tang H, Petrov DA. 2007. Reduced selection leads to accelerated gene loss in Shigella. *Genome Biol*. 8(8):R164.

Hlavacek WS, Faeder JR, Blinov ML, Posner RG, Hucka M, Fontana W. 2006. Rules for modeling signal-transduction systems. *Sci STKE*. 2006(344):re6.

Hurst LD. 2009. Evolutionary genomics and the reach of selection. *J Biol*. 8(2):12.

Huvet M, Toni T, Tan H, Jovanovic G, Engl C, Buck M, Stumpf MPH. 2009. Model-based evolutionary analysis: the natural history of phage-shock stress response. *Biochem Soc Trans*. 37(Pt 4): 762–767.

Joly N, Burrows PC, Engl C, Jovanovic G, Buck M. 2009. A lower-order oligomer form of phage shock protein A (PspA) stably associates with the hexameric AAA(+) transcription activator protein PspF for negative regulation. *J Mol Biol*. 394(4):764–775.

Jones SE, Lloyd LJ, Tan KK, Buck M. 2003. Secretion defects that activate the phage shock response of Escherichia coli. *J Bacteriol*. 185(22):6707–6711.

Jovanovic G, Dworkin J, Model P. 1997. Autogenous control of PspF, a constitutively active enhancer-binding protein of Escherichia coli. *J Bacteriol*. 179(16):5232–5237.

Jovanovic G, Engl C, Buck M. 2009. Physical, functional and conditional interactions between ArcAB and phage shock proteins upon secretin-induced stress in Escherichia coli. *Mol Microbiol*. 74(1):16–28.

Jovanovic G, Engl C, Mayhew AJ, Burrows PC, Buck M. 2010. Properties of the phage shock protein (Psp) regulatory complex that govern signal transduction and induction of the Psp response in Escherichia coli. *Microbiology* 156(Pt 10):2920–2932.

Jovanovic G, Lloyd LJ, Stumpf MPH, Mayhew AJ, Buck M. 2006. Induction and function of the phage shock protein extracytoplasmic stress response in Escherichia coli. *J Biol Chem*. 281(30):21147–21161.

Jovanovic G, Model P. 1997. PspF and IHF bind co-operatively in the psp promoter-regulatory region of Escherichia coli. *Mol Microbiol*. 25(3):473–481.

Jovanovic G, Rakonjac J, Model P. 1999. In vivo and in vitro activities of the Escherichia coli sigma54 transcription activator, PspF, and its DNA-binding mutant, PspFDeltaHTH. *J Mol Biol*. 285(2): 469–483.

Jovanovic G, Weiner L, Model P. 1996. Identification, nucleotide sequence, and characterization of PspF, the transcriptional activator of the Escherichia coli stress-induced psp operon. *J Bacteriol*. 178(7):1936–1945.

Karimpour-Fard A, Leach SM, Gill RT, Hunter LE. 2008. Predicting protein linkages in bacteria: which method is best depends on task. *BMC Bioinformatics*. 9:397.

Kim J-R, Cho K-H. 2006. The multi-step phosphorelay mechanism of unorthodox two-component systems in E. coli realizes ultrasensitivity to stimuli while maintaining robustness to noises. *Comput Biol Chem*. 30(6):438–444.

Kleerebezem M, Crielaard W, Tommassen J. 1996. Involvement of stress protein PspA (phage shock protein A) of Escherichia coli in maintenance of the protonmotive force under stress conditions. *EMBO J*. 15(1):162–171.

Kobayashi R, Suzuki T, Yoshida M. 2007. Escherichia coli phage-shock protein A (PspA) binds to membrane phospholipids and repairs proton leakage of the damaged membranes. *Mol Microbiol*. 66(1):100–109.

Lee SA, Chan CH, Tsai CH, Lai JM, Wang FS, Kao CY, Huang CY. 2008. Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC Bioinformatics*. 9(Suppl 12):S11.

Lewis AC, Saeed R, Deane CM. 2010. Predicting protein-protein interactions in the context of protein evolution. *Mol Biosyst*. 6(1):55–64.

Lloyd LJ, Jones SE, Jovanovic G, Gyaneshwar P, Rolfe MD, Thompson A, Hinton JC, Buck M. 2004. Identification of a new member of the phage shock protein response in Escherichia coli, the phage shock protein G (PspG). *J Biol Chem*. 279(53):55707–55714.

Ludwig W, Schleifer K. 1994. Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. *FEMS Microbiol Rev*. 15:155–173.

Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M. 2004. Genomic analysis of regulatory network dynamics reveals large topological changes.. *Nature*. 431(7006):308–312.

Lynch M. 2007a. The evolution of genetic networks by non-adaptive processes. *Nat Rev Genet*. 8(10):803–813.

Lynch M. 2007b. The origins of genome architecture. Sunderland (MA): Sinauer Associates.

Malpica R, Sandoval GRP, Rodríguez C, Franco B, Georgellis D. 2006. Signaling by the arc two-component system provides a link between the redox state of the quinone pool and gene expression. *Antioxid Redox Signal*. 8(5–6):781–795.

Maxson ME, Darwin AJ. 2006a. Multiple promoters control expression of the Yersinia enterocolitica phage-shock-protein A (pspA) operon. *Microbiology (Reading, Engl)*. 152(Pt 4):1001–1010.

Maxson ME, Darwin AJ. 2006b. PspB and PspC of Yersinia enterocolitica are dual function proteins: regulators and effectors of the phage-shock-protein response.. *Mol Microbiol*. 59(5):1610–1623.

Miller. 1992. A short course in bacterial genetics. Plainview (NY): Cold Spring Harbor Laboratory Press.

Model P, Jovanovic G, Dworkin J. 1997. The Escherichia coli phage-shock-protein (psp) operon. *Mol Microbiol*. 24(2):255–261.

Moreno-Hagelsieb G, Latimer K. 2008. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics*. 24(3):319–324.

Nonaka G, Blankschien M, Herman C, Gross CA, Rhodius VA. 2006. Regulon and promoter analysis of the E. coli heat-shock factor, sigma32, reveals a multifaceted cellular response to heat stress. *Genes Dev*. 20(13):1776–1789.

Organ CL, Janes DE, Meade A, Pagel M. 2009. Genotypic sex determination enabled adaptive radiations of extinct marine reptiles. *Nature* 461(7262):389–392.

Pagel M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of . . .. *Proc Roy Soc B Biol Sci*. 255:37–45.

Paget MSB, Helmann JD. 2003. The sigma70 family of sigma factors. *Genome Biol*. 4(1):203.

Pertea M, Ayanbule K, Smedinghoff M, Salzberg SL. 2009. OperonDB: a comprehensive database of predicted operons in microbial genomes. *Nucleic Acids Res*. 37(Database issue):D479–D482.

Pinney JW, Amoutzias GD, Rattray M, Robertson DL. 2007. Reconstruction of ancestral protein interaction networks for the bZIP transcription factors. *Proc Natl Acad Sci U S A*. 104(51):20449–20453.

Ratmann O, Wiuf C, Pinney J. 2009. From evidence to inference: probing the evolution of protein interaction networks. *HFSP J*. 3:290–306.

Rowley G, Spector M, Kormanec J, Roberts M. 2006. Pushing the envelope: extracytoplasmic stress responses in bacterial pathogens. *Nat Rev Microbiol*. 4(5):383–394.

Thorne T, Stumpf MP. 2007. Generating confidence intervals on biological networks. *BMC Bioinformatics*. 8:467.

Weiner L, Brissette JL, Model P. 1991. Stress-induced expression of the Escherichia coli phage shock protein operon is dependent on sigma 54 and modulated by positive and negative feedback mechanisms. *Genes Dev*. 5(10):1912–1923.

Weiner L, Brissette JL, Ramani N, Model P. 1995. Analysis of the proteins and cis-acting elements regulating the stress-induced phage shock protein operon. *Nucleic Acids Res*. 23(11): 2030–2036.

Wigneshweraraj S, Bose D, Burrows PC, et al. (11 co-authors). 2008. Modus operandi of the bacterial RNA polymerase containing the sigma54 promoter-specificity factor. *Mol Microbiol*. 68(3):538–546.

Xia Y, Franzosa EA, Gerstein MB. 2009. Integrated assessment of genomic correlates of protein evolutionary rate. *PLoS Comput Biol*. 5(6):e1000413.

Yang Z, Bielawski J. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol (Amst)*. 15(12):496–503.

Author/s:
Huvet, M;Toni, T;Sheng, X;Thorne, T;Jovanovic, G;Engl, C;Buck, M;Pinney, JW;Stumpf, MPH

Title:
The evolution of the phage shock protein response system: interplay between protein function, genomic organization, and system function.

Date:
2011-03

Citation:
Huvet, M., Toni, T., Sheng, X., Thorne, T., Jovanovic, G., Engl, C., Buck, M., Pinney, J. W. & Stumpf, M. P. H. (2011). The evolution of the phage shock protein response system: interplay between protein function, genomic organization, and system function.. Mol Biol Evol, 28 (3), pp.1141-1155. https://doi.org/10.1093/molbev/msq301.

Persistent Link:
http://hdl.handle.net/11343/244311